PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : G01L 7/08 — A1

(11) International Publication Number: WO 91/11696

(43) International Publication Date: 8 August 1991 (08.08.91)

(21) International Application Number: PCT/US91/00053

(22) International Filing Date: 2 January 1991 (02.01.91)

(30) Priority data:
474,435 2 February 1990 (02.02.90) US

(71) Applicant: MOTOROLA, INC. [US/US]; 1303 East Algonquin Road, Schaumburg, IL 60196 (US).

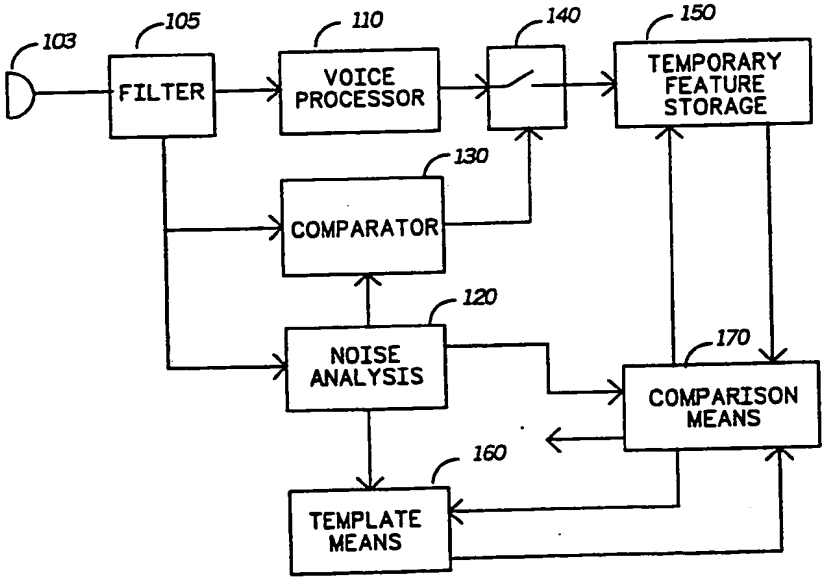
(72) Inventors: ROHANI, Kamyar ; 7050 John T. White #1005, Fort Worth, TX 76120 (US). HARRISON, R., Mark ; 1714 Parkwood Drive, Grapevine, TX 76051 (US).

(74) Agents: PARMELEE, Steven, G. et al.; Motorola, Inc., Intellectual Property Dept., 1303 East Algonquin Road, Schaumburg, IL 60196 (US).

(81) Designated States: AT (European patent), BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), NL (European patent), SE (European patent).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR RECOGNIZING COMMAND WORDS IN NOISY ENVIRONMENTS



(57) Abstract

An input utterance containing a command word to be recognized is processed (110) and features which adequately represent the utterance are determined. Prestored features of a set of reference samples of command words (160) are compared (170) to the features of the input utterance. Recognition of command words in noisy environments is improved by determining the distance between the features of the input utterance and the features of the reference samples and modifying the distance (120) in response to background noise. The reference sample having the minimum distance is selected as the recognized command word.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	ES	Spain	MG	Madagascar
AU	Australia	FI	Finland	ML	Mali
BB	Barbados	FR	France	MN	Mongolia
BE	Belgium	GA	Gabon	MR	Mauritania
BF	Burkina Faso	GB	United Kingdom	MW	Malawi
BG	Bulgaria	GN	Guinea	NL	Netherlands
BJ	Benin	GR	Greece	NO	Norway
BR	Brazil	HU	Hungary	PL	Poland
CA	Canada	IT	Italy	RO	Romania
CF	Central African Republic	JP	Japan	SD	Sudan
CG	Congo	KP	Democratic People's Republic of Korea	SE	Sweden
CH	Switzerland			SN	Senegal
CI	Côte d'Ivoire	KR	Republic of Korea	SU	Soviet Union
CM	Cameroon	LI	Liechtenstein	TD	Chad
CS	Czechoslovakia	LK	Sri Lanka	TG	Togo
DE	Germany	LU	Luxembourg	US	United States of America
DK	Denmark	MC	Monaco		

METHOD AND APPARATUS FOR RECOGNIZING COMMAND WORDS IN NOISY ENVIRONMENTS

Technical Field

This invention relates generally to the field of word recognizers and in particular to those word recognizers which are capable of recognizing command words in noisy environments.

Background

Traditionally, the interaction between humans and devices has been achieved by some form of manual interaction, such as activating a switch or pushing a button. However, in many instances it may be advantageous or even necessary to interface with the device by means of a voice command. For example, a policeman in a police car may activate numerous functions, such as turning on the siren, by simply uttering an appropriate utterance which contains a word command. A word recognizer after receiving and processing the utterance, recognizes the word command and effectuates the desired function.

Generally, a word recognizer recognizes the word command by extracting features which adequately represent the utterance, and making a decision as to whether these features meet a particular criteria. These criteria may comprise correspondence to a set of pre-stored features representing the command words to be recognized.

The word recognizer may be speaker dependent or speaker independent. A speaker independent word recognizer is

designed to recognize the commands of potentially any number of users regardless of the differences in speech patterns, accents, and other variations in spoken words. However, the speaker independent word recognizer requires significantly sophisticated processing capability and hence has been constrained to recognizing a limited number of command words.

A speaker dependent word recognizer is designed to recognize the command words of limited number of users by comparing the utterance to prestored voice templates which contain the voice features of those users. Therefore, it is necessary to train the word recognizer to recognize the voice features of each individual user. Training is commonly understood to be a process by which the individual users repeats a predetermined set of word commands for a sufficient number of times so that an acceptable number of their voice features are extracted and stored as reference features.

One of the important characteristics of a word recognizer is its capability to accurately recognize a word command under various noise conditions. Typically, the word recognizers provides error rates of less than 1% in quiet environments. However, the error rate may be degraded by as much as 40% in environments where there is a 20 db peak signal-to-noise ratio (SNR). One of the factors contributing to poor noise performance is the difference between the training condition under which the reference features are derived and the operating condition under which the utterance features are derived. Accordingly, due to this difference, comparison of the reference features and input utterance features may produce substantially erroneous results.

Many word recognizers incorporate noise compensation techniques in means utilized to derive the reference features. In one such word recognizer, a background noise estimator provides the ambient noise characteristics, and the prestored reference features are temporarily modified according to the characteristics of the ambient noise. The modified reference features and the input utterance features are then compared to each other, and the reference sample having features with the

closest similarity to the features of the input utterance is declared as the recognized word.

In another type of word recognizer, the features of the input utterance are represented by the amount of energy contained within predetermined number of frequency bands. This technique is known as the filter banks method. In the word recognizer utilizing this technique the noise compensation is achieved by determining the back ground noise energy at every frequency band and subtracting it from the energy at the corresponding frequency band of the input utterance. The resulting features are then compared to the corresponding reference features, and again the reference sample having most similar features to the features of the input utterance is declared the recognized word. However, this type of system suffers an inherent draw back in that the number of predetermined frequency band is critical to the proper operation of the word recognizer. That is, dividing the voice spectrum into a high number of frequency band causes degradation in recognition accuracy of high pitched voices, and dividing the voice spectrum into a low number of frequency bands causes smearing effect on the voice signal.

Other means of noise compensation in speech recognition utilize noise reduction techniques, wherein the signal to noise ratio is increased using various filtering techniques. However, practical improvements in SNR typically fall short of achieving a substantial accuracy in recognizing word commands. Another method of noise compensation for a system utilized in a severe noise environment is to train the system in a comparable noise environment. However, certain type of noise, such as acoustical background noise, are time variant in nature. Accordingly it is not possible to predict or otherwise reproduce, during training, the actual time variant noise which will exist during a subsequent speech recognition mode.

Summary of the Invention

Accordingly, it is an object of the present invention to provide a word recognizer apparatus capable of accurately recognizing command words under various noise conditions.

Briefly, the word recognizer of the invention comprises a voice processing means for receiving an input utterance and determining features which adequately represent the utterance. A template means provides the pre-stored features of a set of reference samples which represent the recognizable command words. A noise analysis means determines ambient noise characteristics. A comparison means determines the distance between the features of the utterance and the reference samples. The comparison means is responsive to the ambient noise characteristics for modifying the determined distance. The word recognizer apparatus include means for determining the minimum distance and selecting the reference sample based thereon.

Brief Description of the Drawings

Figure 1 shows a block diagram of the word recognizer of the invention.

Figure 2, shows a block diagram of the voice processor shown in Figure. 1.

Figure 3, is the flow chart for extracting CSM features of an input utterance.

Figure 4, shows the block diagram of the noise analyzer of FIG.1.

Figure 5, shows a portion of the word recognizer of the invention which includes the block diagram of the template means of Figure 1.

Figure 6, shows the graph of the power distribution of the reference sample and the input utterance for a command word.

Figure 7, shows a portion of the word recognizer of the invention which includes the block diagram of the comparison means of Figure 1.

Figure 8, is the flow chart of the steps taken according to the invention to recognize the word command in noisy environments.

Detailed Description of the Preferred Embodiment

Referring to FIG. 1, a the block diagrams shown of a word recognizer 100 which utilizes the principals of the present invention for recognizing word commands. The word recognizer 100 comprise an isolated word recognizer which is capable of recognizing more than one spoken word commands having a pause therebetween. The word recognizer 100 includes a voice processor 110 for processing an input utterance containing one or more word commands. The input utterance is received through a microphone 103 which produces a voice signal representing the input utterance. A well known audio filter 105 is used to limit the frequency spectrum of the input utterance to a predetermined range. In the preferred embodiment of the invention, the range of the audio filter 105 is confined to a range of 200 Hz to 3200 Hz. The voice processor 110 divides the input utterance in to frames of predetermined duration. The voice processor 110 provides, in each frame, those features of the input utterance which adequately characterize the input utterance. The detailed process by which these features are produced is described later. These features comprise frequency components and corresponding amplitudes as well as the power of the input utterance in each frame. A background noise analyzer 120 provides the characteristics of the ambient noise. These characteristics comprise signal to noise ratio in the frequency spectrum and the level of the ambient noise floor. Because the word recognizer 100 is an isolated word recognizer, the beginning and the end of the input utterance must be determined. In the preferred embodiment of the invention, this determination is made by comparing the power of the input utterance to the power of the ambient noise floor. When the power of the input utterance exceeds the ambient noise floor a comparator 130 closes a switch 140, thereby allowing the features of the input utterance to

be stored in a temporary feature storage means 150. When the power of the input utterance falls below the ambient noise floor the switch 140 is opened preventing features from being stored in the storage means 150. Accordingly, the end points of the input utterance are determined by comparing the ambient noise floor to the power of the input utterance. A template means 160 provides the features of a set of prestored reference samples. The features of the prestored reference samples are generated, during training, utilizing the same process as that which provides the features of the input utterance. As subsequently described herein, the template means 160 aligns the end points of the reference sample with the end points of the input utterance. A comparison means 170 primarily comprising a microcomputer/controller provides the distance between the features of the input utterance and the reference samples. The detail of the process by which the distance between the features of the input utterance and the reference sample are produced is described later. The comparison means 170 then selects the reference sample having the minimum distance with the features of the input utterance and based thereon declares the word command. Noise compensation in the word recognizer of the invention is achieved by eliminating or modifying the distance between the features of the input utterance and the features of the reference sample having noise characteristics above a predetermined threshold.

Referring to FIG. 2, the block diagram of the voice processor 110 comprises an A/D converter 102 which samples the voice signals provided by microphone 103 of FIG. 1 at a suitable sampling rate,such as 8000 samples per second. A frame buffer 104 buffers the sampled signal and provides frames which consist of a predetermined number of consecutive voice samples. The framing technique utilized by the frame buffer 104 is well known in the art, and the frames provided by the preferred embodiment of the invention comprise 160 samples which correspond to a frame duration of 20 msec. It may be appreciated that depending on the duration of each input utterance a variable

number of frames (designated as N) may be generated by the frame buffer 104.

The features characterising each frame utterance may be parametric or discrete. The discrete features of the utterance frames may be provided by such known techniques as the filter banks method. The embodiment of the present invention utilizes a technique which provides the parametric features of the utterance frame. The parametric features of the utterance may be provided by such known techniques as linear predictive analysis (LPC) or composite sinusoidal modeling (CSM). In the preferred embodiment of the invention, the features of the utterance frames are provided utilizing conventional CSM analysis techniques as described in S. Sagayama and F. Ikatura, "Duality Theory of Composite Sinusoidal Modelling and Linear Prediction", ICASSP '86 Proceedings, vol 3, pp. 1261-1264, the disclosure of which is hereby incorporated by reference. The purpose of CSM analysis is to determine a set of CSM features which adequately characterize the frame utterance. The CSM features comprise CSM frequencies { f_j } and amplitudes { m_j } which correspond thereto. The number of CSM features (designated as M) of each frame of the input utterance is related to the frequency range of the utterance. In utterances confined to a range of 200 Hz to 3200 Hz in frequency spectrum there usually exists four formant resonant frequencies below 3200 Hz. Thus, it is usually sufficient to utilize 4 CSM frequencies and amplitudes to characterize the input utterance frames. Therefore, in the preferred embodiment of the invention, the number of features (designated as M) is equal to 4. A feature extractor 106 executes a feature extraction process utilizing conventional CSM techniques which as shown in the flow chart of FIG.3.

According to FIG. 3, the CSM extractor 106, at block 310, applies the input utterance features and computes the autocorrelation of the frame utterances at block 320. The term of the interpolative correlation is then computed, block 330. At block 340, the CSM extractor 106 solves a Hankel matrix for providing the coefficients of a polynomial:

$$P_n+.....+P_1\ x^{n-1}+x^n=0 \qquad (1).$$

Then the real roots { x_i } of the equation (1) are provided, block 350. The amplitude CSM features, { m_i }, are provided by the matrix shown in a block 360, and the frequency CSM features, { f_i }, are provided by the following equation:

$$\{ f_i \} = \mathrm{Cos}^{-1} x_i \qquad (2).$$

In addition to amplitude and frequency CSM features, the feature extractor 106 also provides the power content of the frame input utterance frame derived from the following equation:

$$P(n) = (1/N) \sum_{i=1}^{N} \{T(i)\}^2 \qquad (3).$$

Accordingly, the features of the input utterance for each frame may be represented by a composite vector:

$$T(n) = \{ m_1^n, m_2^n, ..., m_M^n, f_1^n, f_2^n, ..., f_M^n, P_{(n)} \} \quad (4)$$

and the entire utterance may be represented by

$$\{T_1, T_2, ..., T_N\} \qquad (5).$$

One of ordinary skill in the art may appreciate that the voice processor 110 described in FIG. 2 and in FIG. 3 may be implemented by means of any suitable digital signal processor (DSP), such as 56000 series family of DSPs manufactured by Motorola, Inc.

Referring to FIG. 4, the block diagram of a well known noise analyzer 120 is shown. The noise analyzer 120, continually monitors the background noise and provides characteristics thereof. The noise analyzer 120 includes a noise processing means 122 for producing the noise powers of the desired frequency spectrum. The noise processing means 122 utilizes well known analysis techniques, such as Fast Fourier Transformation analysis, to provide noise power at desired CSM frequencies. The noise processor 122 also receives the corresponding CSM amplitudes of the input utterance frames and produces the signal to noise ratios SNR (f) at the CSM frequencies. Additionally the noise analyzer 120 includes a well known noise averaging means 124 which provides the power at

noise floor Rn. The techniques for providing ambient noise floor is well known in the art.

Referring to FIG. 5, the block diagram of the template means 160, which in the preferred embodiment of the invention, operates under the control of the comparison means 170, is shown. The template storage means 162 stores the features of a set of reference samples representing word commands recognizable by the word recognizer 100. These reference features have been obtained during a training process. During the training process, a user repeats each of the desired word commands to be recognized a number of times. Preferably, the training of the word recognizer is performed in a quiet environment. The features of the user voice are extracted and stored in the template storage means 162 as the reference samples. During training, the utterances are processed identically to the processing of the input utterance. In fact, the voice processor 110 is used to generate the reference sample features during training of the word recognizer 100. It may be appreciated that the number of reference sample frames (designated as J) may be different from the number of the corresponding input utterance frames N. It should be noted that the powers of each frame as derived from equation (3) are also included in the features of the reference sample. Accordingly, the features of the each reference sample may be stored in the template storage means 162 as vectors:

$$R(j) = \{ m_1^j, m_2^j, \ldots, m_M^j, f_1^j, f_2^j, \ldots, f_M^j, P_{(j)} \}, \quad (6)$$

and the all of the recognizable reference samples are represented as:

$$W_1 = \{R^1_1, \ldots, R^1_{j(1)}\} \quad (7)$$

..

$$\ddot{W}_K = \{R^K_1, \ldots, R^K_{j(K)}\} \text{ where} \quad (8)$$

J(1),...,J(K) = number of frames in the reference samples, and

K= number of reference samples.

In operation and under the control of comparison means 170, each of these reference samples are selected and compared to the input utterance. In order to achieve an effective comparison, the end points of the reference sample under comparison and the input utterance must be aligned. However, because the features of the reference sample are generated in a quiet environment and these same features are used for comparison under noisy conditions, the end points of the reference sample and the input utterance may become misaligned. In the preferred embodiment of the invention, an end point aligner 164 is included in the template means 160 to alleviate end point misalignments. FIG. 6 shows in time domain the power contour 610 of a reference sample for a word command. It may be appreciated that the power contour 610 of the reference sample can actually be represented by a number of discrete powers corresponding to each frame. However, for the sake of simplicity and ease of understanding the contour of the power distribution of the reference sample is shown as a solid line 610. Similarly, the power contour of an input utterance substantially corresponding to that of the reference sample is shown by a dotted line 620. As shown, the end points of the reference sample in quiet background and the input utterance in noisy environments are separated from each other by the ambient noise floor power R (n). It may be appreciated that if the end points of the reference sample are readjusted by a number of frames such that the subsequent frames have powers above the noise floor power, the end points of the reference sample and the input utterance may be realigned. Therefore, the noise floor power Rn provided by noise analyzer 120 constitutes a threshold by which the end points of the reference sample are readjusted. Referring back to FIG.5, the end point aligner 164 skips those candidate endpoints whose power are below the noise power. One of ordinary skill in the art appreciates that the the end point aligner 164 may be implemented by means of any suitable

microcomputer or DSP executing a suitable program for achieving the intended purpose thereof.

Referring to FIG. 7, the comparison means 170 comprise a well known a microcomputer/controller, such as the 68000 family of microcomputers manufactured by Motorola, Inc. The comparison means 170, among other things, includes a controller 172, a computer 174, a RAM 176 and a ROM 178. The controller 172 performs several functions which include controlling the operation of the comparison means 170 and the template means 160 as well as interacting with the temporary storage means 150 and noise analyzer 120. The computer 174 performs the computational functions of the comparison means 170. The RAM 176 provides a temporary information storage for the computer 174 and the controller 172. The program containing the operational steps of the computer 174 and the controller 172 is stored in the ROM 178.

The operation of the comparison means is described in conjunction with the flow chart shown in FIG 8. At block 810, the controller 172 receives the features of the input utterance from the temporary storage means 150. At block 820, the features of the first reference sample after endpoint alignment are received from the template means 160. At block 830, the computer 174 determines the distance between the features of the reference sample and the input utterance. In the preferred embodiment of the invention, only the frequency features of the frames of the utterance and the reference sample are utilized for computing the distance. The determined distance is called a local distance metric and is computed from the following equation:

$$d(n,j) = \sum_{i=1}^{M} (T(i,n)-R(i,j))^2 \qquad (9)$$

where

$1 <= i <= M$ is the Index of composite sinusoidal features,

$1 <= n <= N$ is the time index of utterance,

$1 <= j <= J$ is the time index of the reference sample,

T (i,n) represents the i th composite sinusoidal frequency in the n th frame of said utterance,

R (i, j) represents the i th composite sinusoidal frequency in the j th frame of said reference sample.

After the local distance for each frequency feature of every frame is calculated, the local distant metric d is modified by a function W(i, n) of the signal to noise ratio SNR(f) provided by the noise analyzer 120. The function W(i,n) may be defined as:

$$W(i,n)=F[SNR](f)\} \qquad (10).$$

Therefore the modified local distance may be defined as:

$$d=\sum_{i=1}^{M} (T(i,n)-R(i,j))^2 * W(i,n) / K, \quad (11)$$

where K is the normalization constant defined by:

$$K=\sum_{n=1}^{N} \sum_{i=1}^{M} W(i,n).$$

In the preferred embodiment of the invention W(i,n) comprises a discrete function defined by:

$$W(i,n)=\begin{cases} 1 & \text{If SNR (f) > N.T.} \\ 0 & \text{If otherwise.} \end{cases} \qquad (12)$$

where N.T= signal to noise ratio threshold. Accordingly, the ith frequency features of the nth frame is eliminated, if the SNR(f) at that frequency is below the SNR threshold. The W(i,n) may comprise a continuously differentiable limiting function, such as well known sigmoidal or hyprobolic tangent functions . It may be appreciated that for each frame of the input utterance there is total of at most J local distances. The legal local distance minimum of each input utterance frame are added to subsequent local distances. An accumulated distance is thus determined for each reference sample frame, block 840. One of ordinary skill in the art may appreciate that under predetermined boundary and

continuity conditions a minimum distance may be obtained utilizing well known dynamic time warping techniques. One such technique is described in ITAKURA, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE proceedings on Acoustics, Speech, & Signal Processing, vol ASSP-23, No.1, pp. 67-72, February, 1975 which is hereby incorporated by reference. At block 850, the minimum distance utilizing such a technique is computed and stored. In block 860, a decision is made to determine whether more reference samples are to be processed. After comparing all of the stored reference samples, block 870, the reference sample having the minimum distance is selected. In block 880, the command word contained in the input utterance is recognized based on a decision on the selected reference sample. The decision also takes into consideration a predetermined criteria before the recognized command word is declared. Such criteria may comprise a threshold minimum distance below which the recognized word is valid. This predetermined criteria prevents declaring an invalid input utterance, which produces a minimum distance, the recognized word command.

As described, during the recognition process, the local distances between the features of the input utterance and the reference sample are relied upon in recognizing the command word. The local distances are modified as a function of the signal to noise ratio. Accordingly, the accuracy of the word recognizer under severe noise conditions is improved by eliminating or lessening the contribution of those local distances which have an undesirable noise characteristic.

Claims

1. An apparatus for recognizing command words in an utterance comprising:

voice processing means for determining features representing said utterance;

template means for providing features of a plurality of reference samples representing command words to be recognized by said apparatus;

noise analysis means for determining ambient noise characteristics;

comparison means including means for determining the distances between features of said utterance and each reference sample being responsive to said ambient noise characteristics for modifying the determined distance,

decision means including means for determining the minimum modified distance, and mean for recognizing the command word based on said minimum distance.

2. The apparatus of claim 1, wherein said features of said utterance and said reference samples are parametric features.

3. The apparatus of claim 2, wherein said template means comprises means for aligning said utterance and said reference sample end points.

4. The apparatus of claim 2, wherein said parametric features include frequencies and corresponding amplitudes for said frequencies.

5. The apparatus of claim 4, wherein said ambient noise characteristics include signal to noise ratio at said frequencies.

6. The apparatus of claim 5, wherein said comparison means is responsive to said ambient noise characteristics for modifying said distance as a function of said signal to noise ratio.

7. The apparatus of claim 6, wherein said distance is modified when said signal to noise ratio exceeds a predetermined threshold.

8. The apparatus of claim 6, wherein said function for modifying said distance comprise a continuously differentiable limiting function.

9. The apparatus of claim 6, wherein said distance comprises a local distance metric defined by:

$$d=\sum_{i=1}^{M} (T(i,n)-R(i, j))^2$$

where

$1 <= i <= M$ is the Index of composite sinusoidal features,

$1 <= n <= N$ is the time index of utterance,

$1 <= j <= J$ is the time index of the reference sample,

T (i,n) represents the i th composite sinusoidal features in the n th frame of said utterance,

R (i, j) represents the i th composite sinusoidal features in the j th frame of said reference sample

10. The apparatus of claim 9, wherein said recognition means include means for determining the minimum distance by utilizing a dynamic time warping technique.

11. A method for recognizing command words in an utterance comprising:

a) determining features of said utterance;

b) providing features of a plurality of reference samples representing command words to be recognized;

c) determining characteristics of ambient noise;

d) determining the distance between features of said utterance and features of each of said plurality of reference samples;

e) modifying the distance between features of said utterance and features of each of said plurality of reference samples in response to said characteristics of ambient noise;

f) determining the minimum of the modified distance; and

g) recognizing the command word based on the minimum distance.

12. The method of claim 11, wherein said step (a) comprises determining parametric features of said utterance and step (b) comprises providing parametric features of reference samples.

13. The apparatus of method 12, wherein said steps (b) includes aligning of said reference sample and said utterance end points.

14. The method of claim 12, wherein steps (a) includes determining frequencies and corresponding amplitudes for said frequencies and step (b) includes providing composite sinusoidal frequencies and amplitudes at said frequencies.

15. The method of claim 14, wherein said step (c) includes determining signal to noise ratio at said frequencies.

16. The method of claim 15, wherein said step (e) comprise modifying the distance between features of said utterance and features of each of said plurality of reference samples as a function of signal to noise ratio.

17. The method of claim 16, wherein said step (e) comprise modifying the distance between features of said utterance and features of each of said plurality of reference samples when the signal to noise ratio exceeds a predetermined ratio.

18. The method of claim 16, wherein said step (e) comprises modifying the distance between features of said utterance and features of each of said plurality of reference samples when the signal to noise ratio exceeds a predetermined ratio.

NM19. The of method of claim 25, wherein said step (d) of determining the distance between composite sinusoidal features of said utterance and composite sinusoidal features of each of said plurality of reference samples is derived from a local distance metric function defined by:

$$d=\sum_{i=1}^{M} (T(i,n)-R(i, j))^2$$

where

$1 <= i <= M$ is the Index of composite sinusoidal features,

$1 <= n <= N$ is the time index of utterance,

$1 <= j <= J$ is the time index of the reference sample,

T (i,n) represents the i th composite sinusoidal features in the n th frame of said utterance,

R (i, j) represents the i th composite sinusoidal features in the j th frame of said reference sample

20. The method of claim 16, wherein said step (f) comprises determining the minimum of the modified distance by utilizing a dynamic time warping technique.

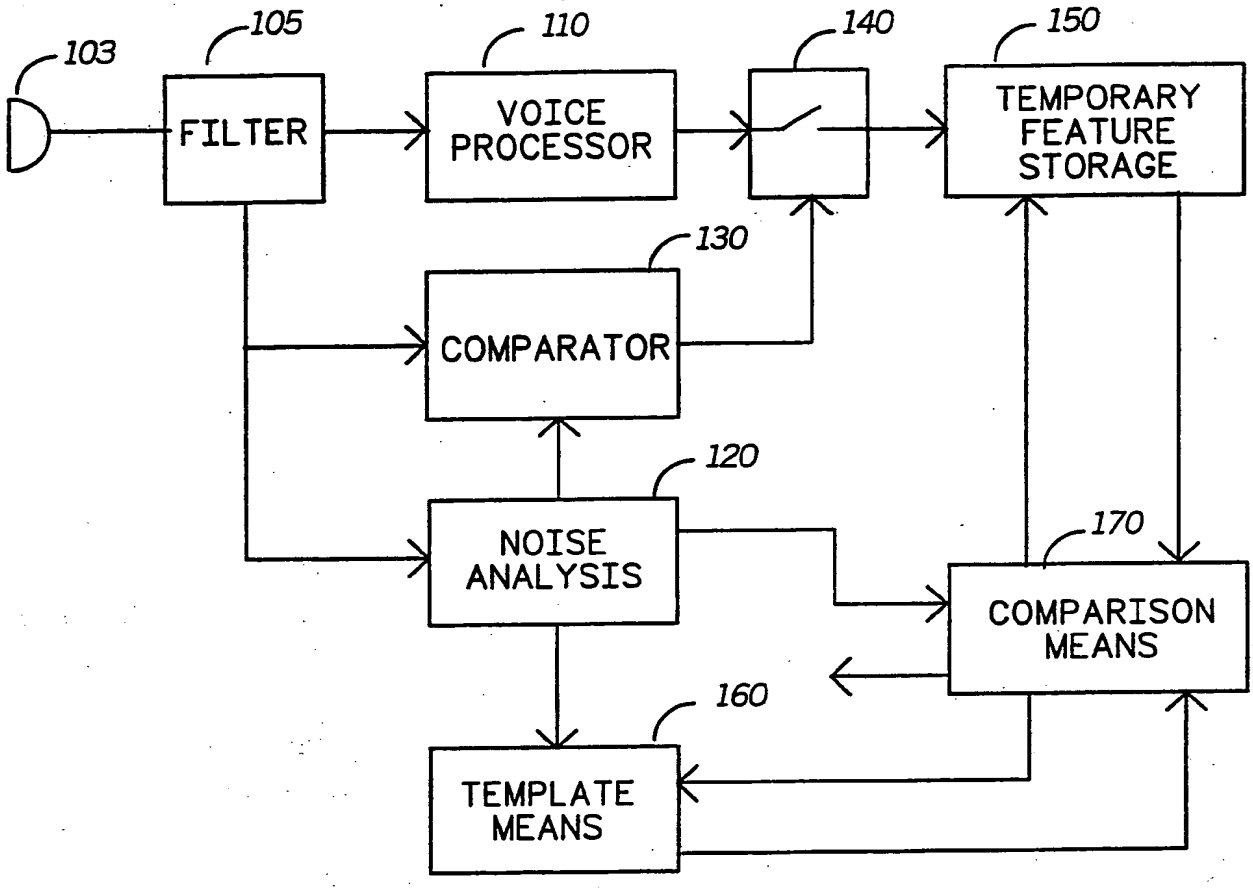

103
105
FILTER
110
VOICE PROCESSOR
140
150
TEMPORARY FEATURE STORAGE
130
COMPARATOR
120
NOISE ANALYSIS
170
COMPARISON MEANS
160
TEMPLATE MEANS

FIG.1

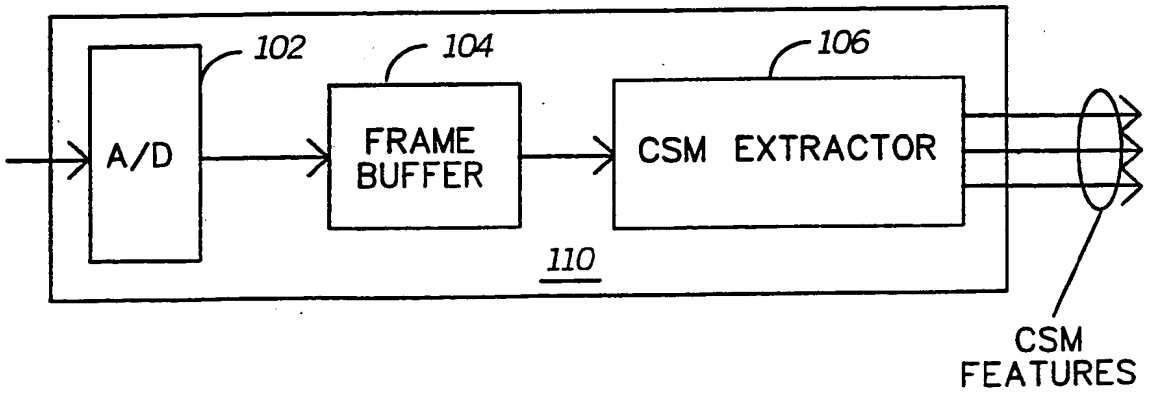

102
A/D
104
FRAME BUFFER
106
CSM EXTRACTOR
110
CSM FEATURES

FIG.2

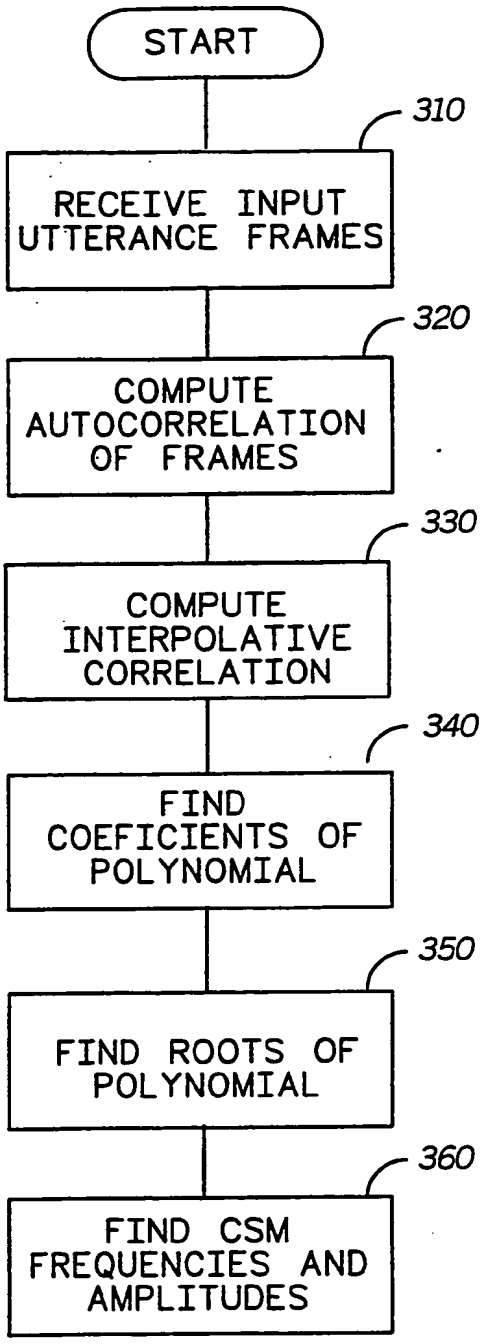

START
310
RECEIVE INPUT UTTERANCE FRAMES
320
COMPUTE AUTOCORRELATION OF FRAMES
330
COMPUTE INTERPOLATIVE CORRELATION
340
FIND COEFICIENTS OF POLYNOMIAL
350
FIND ROOTS OF POLYNOMIAL
360
FIND CSM FREQUENCIES AND AMPLITUDES

FIG.3

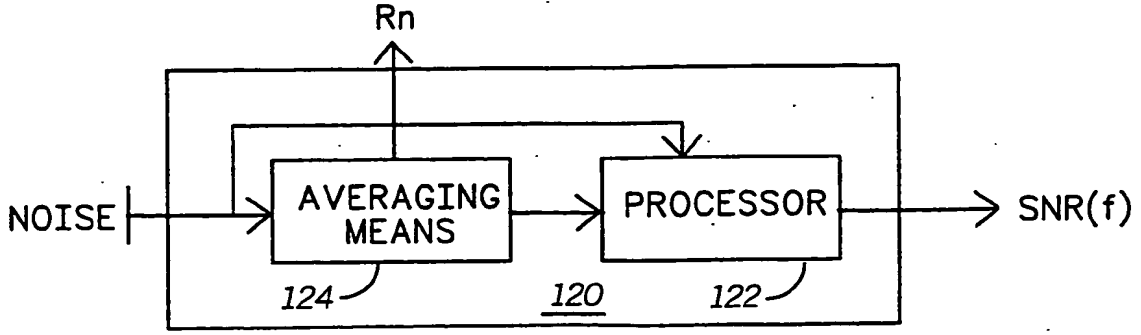

Rn
NOISE
AVERAGING MEANS
PROCESSOR
SNR(f)
124
120
122

FIG.4

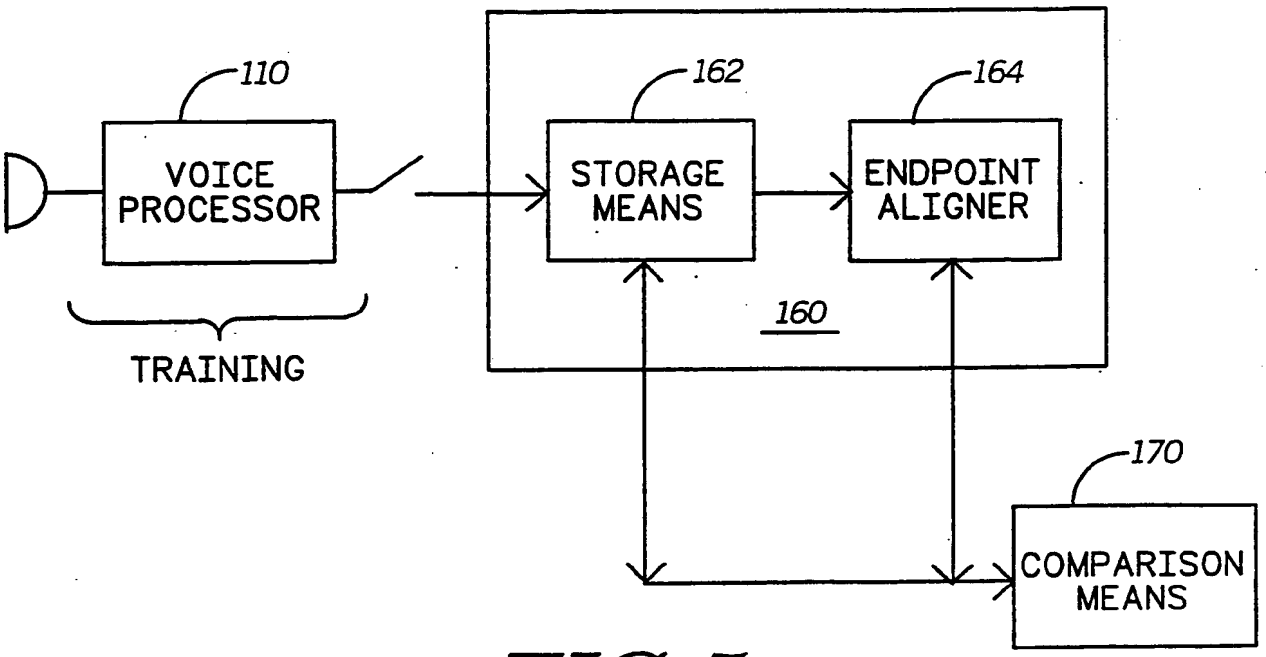

110
VOICE PROCESSOR
TRAINING
162
STORAGE MEANS
164
ENDPOINT ALIGNER
160
170
COMPARISON MEANS

FIG. 5

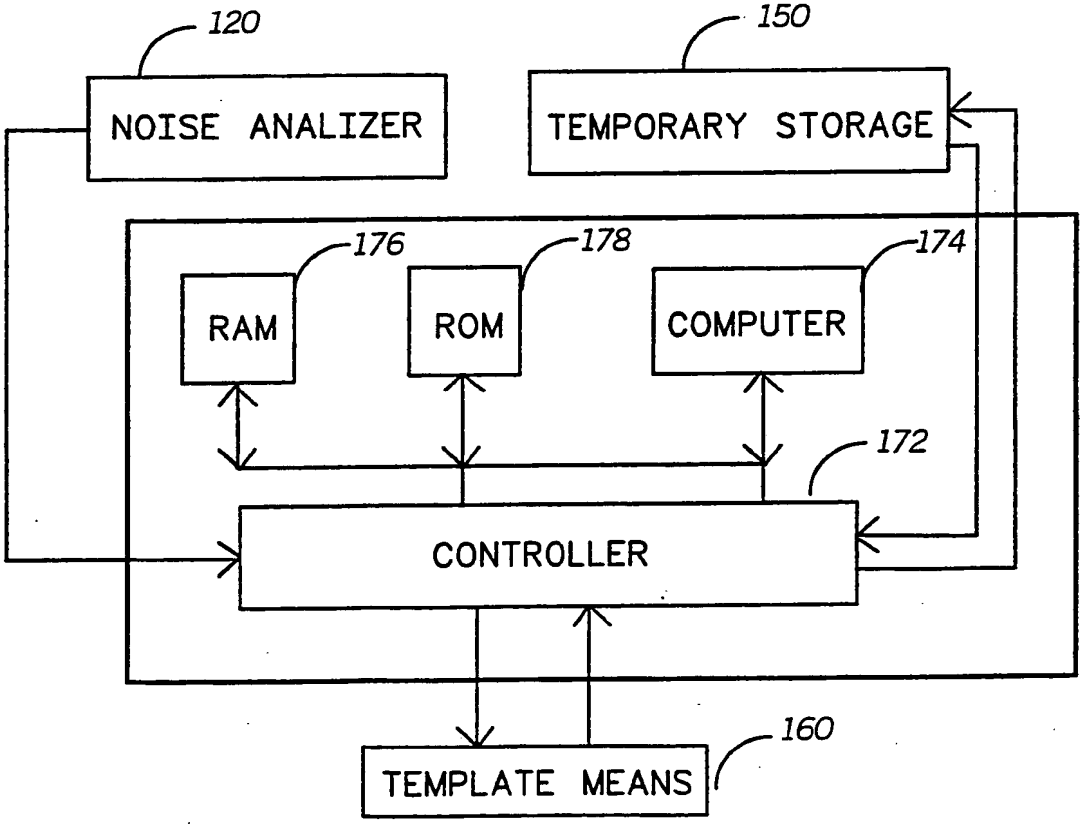

120
NOISE ANALIZER
150
TEMPORARY STORAGE
176
RAM
178
ROM
174
COMPUTER
172
CONTROLLER
TEMPLATE MEANS
160

FIG. 7

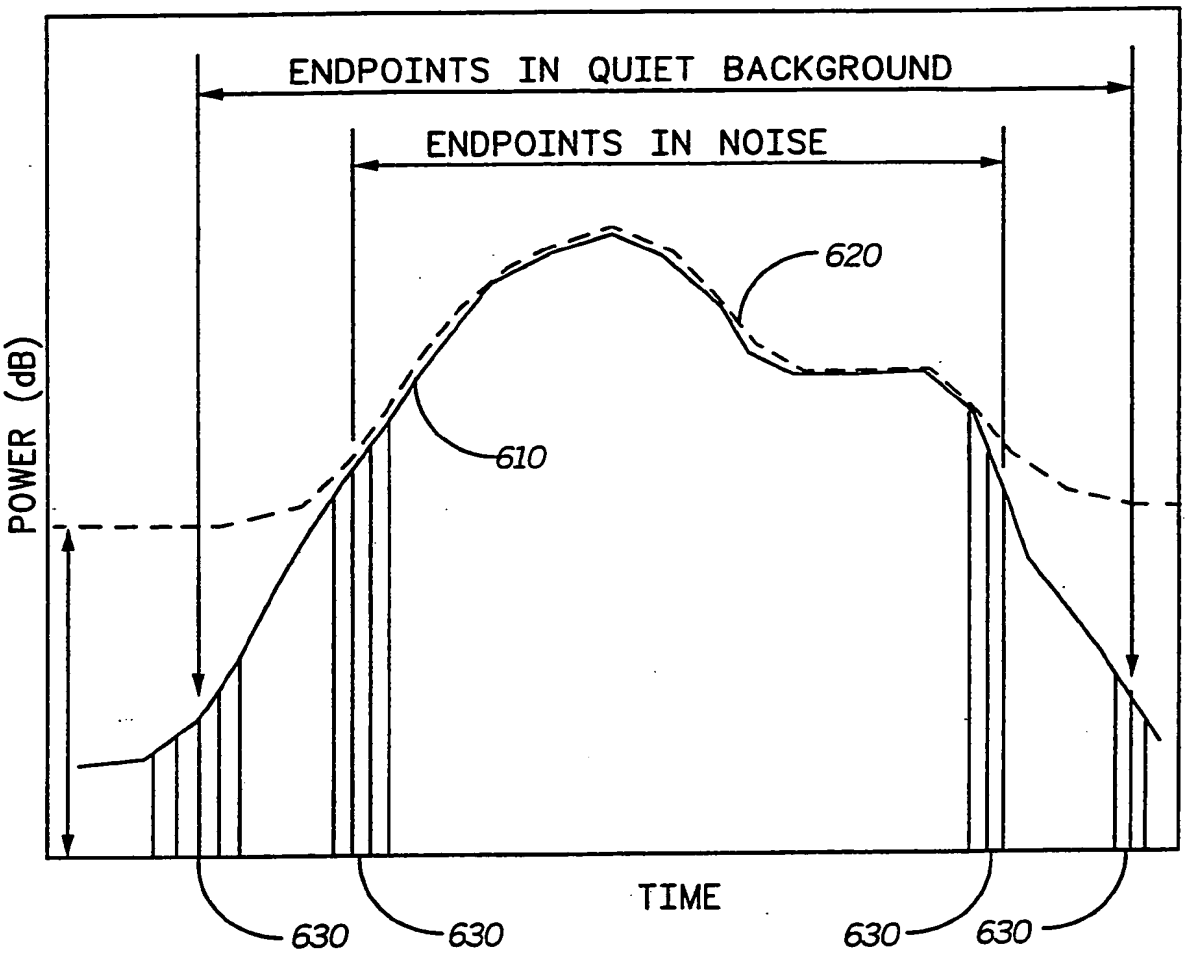

ENDPOINTS IN QUIET BACKGROUND
ENDPOINTS IN NOISE
620
610
POWER (dB)
TIME
630
630
630
630

FIG.6

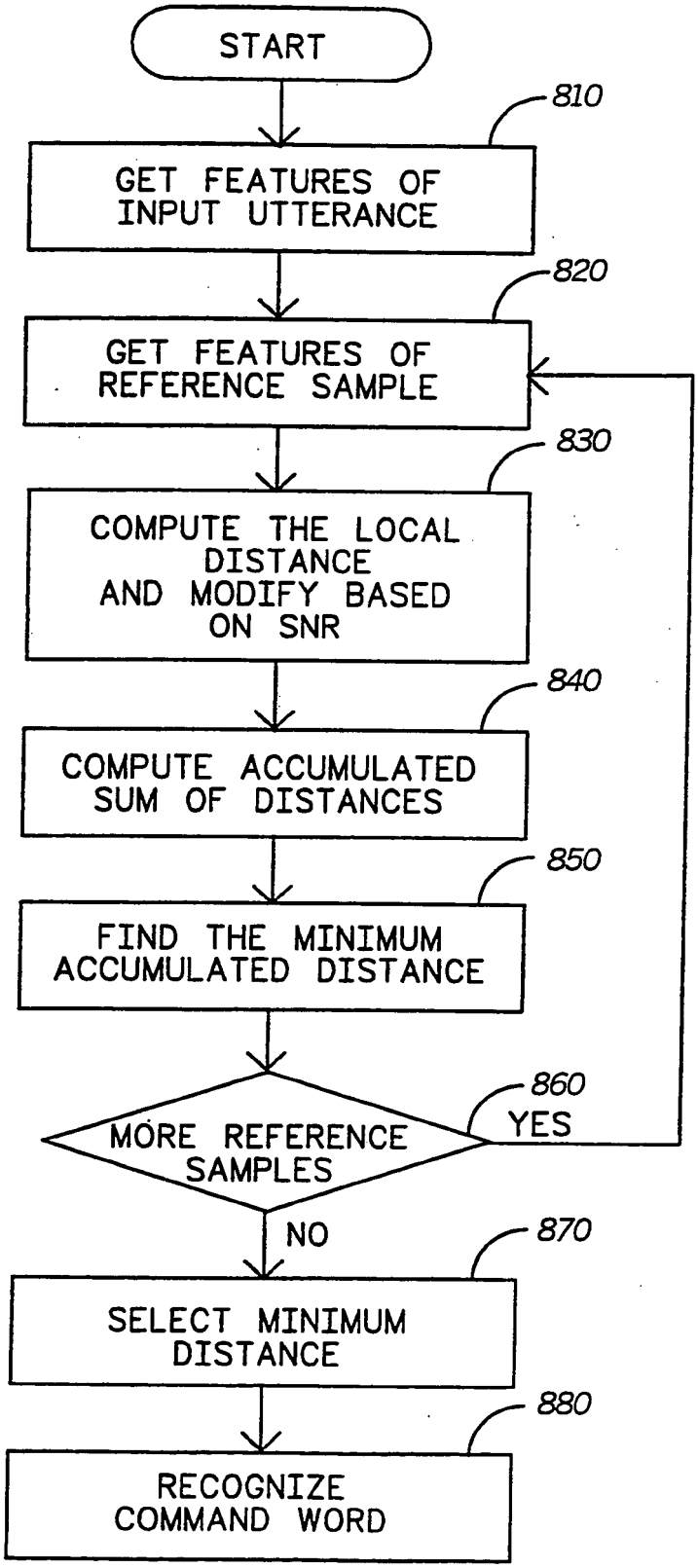

START
810
GET FEATURES OF INPUT UTTERANCE
820
GET FEATURES OF REFERENCE SAMPLE
830
COMPUTE THE LOCAL DISTANCE AND MODIFY BASED ON SNR
840
COMPUTE ACCUMULATED SUM OF DISTANCES
850
FIND THE MINIMUM ACCUMULATED DISTANCE
860
MORE REFERENCE SAMPLES
YES
NO
870
SELECT MINIMUM DISTANCE
880
RECOGNIZE COMMAND WORD

FIG. 8

INTERNATIONAL SEARCH REPORT

International Application No PCT/US91/00053

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) [3]

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC (5): G01L 7/08
U.S. CL: 381/43

II. FIELDS SEARCHED

Minimum Documentation Searched [4]

Classification System	Classification Symbols
U.S.	381/41, 46, 110 364/513.5 367/198

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched [5]

III. DOCUMENTS CONSIDERED TO BE RELEVANT [14]

Category *	Citation of Document, [16] with indication, where appropriate, of the relevant passages [17]	Relevant to Claim No. [19]
Y	US,A, 4,829,578 (Roberts), 9 May 1989.	1-18, 20
Y	US,A, 4,852,181 (Morito et al), 25 July 1989.	1-18, 20
Y	US,A, 4,897,878 (Boll et al), 30 January 1990.	1-18, 20
Y,P	US,A, 4,918,732 (Gerson et al), 17 April 1990.	1-18, 20
Y,P	US,A, 4,933,973 (Porter) 12 June 1990.	1-18, 20
Y	UK,A, 2,137,791 (Bridle et al.), 10 October 1984.	1-18, 20

* Special categories of cited documents: [15]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search [2]

15 March 1991

Date of Mailing of this International Search Report [2]

18 APR 1991

International Searching Authority [1]

ISA/US

Signature of Authorized Officer [20]

John A. Merecki

FURTHER INFORMATION CONTINUED FROM THE SECOND SHEET

V. ☐ OBSERVATIONS WHERE CERTAIN CLAIMS WERE FOUND UNSEARCHABLE [1]

This international search report has not been established in respect of certain claims under Article 17(2) (a) for the following reasons:

1. ☐ Claim numbers . . . . . , because they relate to subject matter [1] not required to be searched by this Authority, namely:

2. ☐ Claim numbers . . . . . , because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out [1], specifically:

3. ☒ Claim numbers 19, because they are dependent claims not drafted in accordance with the second and third sentences of PCT Rule 6.4(a).

VI. ☐ OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING [2]

This International Searching Authority found multiple inventions in this international application as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims of the international application.

2. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims of the international application for which fees were paid, specifically claims:

3. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claim numbers:

4. ☐ As all searchable claims could be searched without effort justifying an additional fee, the International Searching Authority did not invite payment of any additional fee.

Remark on Protest

☐ The additional search fees were accompanied by applicant's protest.

☐ No protest accompanied the payment of additional search fees.